MODELLING SEMANTIC CONTEXT FOR NOVELTY DETECTION IN WILDLIFE SCENES

Suet-Peng Yong, Jeremiah D. Deng, and Martin K. Purvis

Dept. of Information Science, University of Otago, PO Box 56, Dunedin, New Zealand
Email: {spyong,ddeng,mpurvis}@infoscience.otago.ac.nz

ABSTRACT

Novelty detection is an important functionality that has found many applications in information retrieval and processing. In this paper we propose a novel framework that deals with novelty detection for multiple-scene image sets. Working with wildlife image data, the framework starts with image segmentation, followed by feature extraction and classification of the image blocks extracted from image segments. The labelled image blocks are then scanned through to generate a co-occurrence matrix of object labels, representing the semantic context within the scene. The semantic co-occurrence matrices then undergo binarization and principal component analysis for dimension reduction, forming the basis for constructing one-class models for each scene category. An algorithm for outlier detection that employs multiple one-class models is proposed. An advantage of our approach is that it can be used for scene classification and novelty detection at the same time. Our experiments show that the proposed approach algorithm gives favourable performance for the task of detecting novel wildlife scenes, and binarization of the label co-occurrence matrices helps to significantly increase the robustness in dealing with the variation of scene statistics.

Keywords— context, co-occurrence matrix, semantics, novel image, multi-class

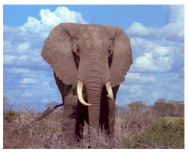
1. INTRODUCTION

Something out of the conventional context in our perception is considered as either abnormal or interesting to us. According to psychological studies, novelty seems intuitively tied to interest [Sil06], but short-term novelty can also be identified with something that contrasts with recent experience [Ber60]. The modelling of this novelty detection mechanism, together with image analysis techniques, can be useful in a number of applications. For instance on a social network such as Flickr or Facebook, images can be automatically analyzed, and abnormal or inappropriate content can be detected and then either removed or hidden from certain users or groups. Similarly, when browsing a large photo collection, it is desirable to use novelty detection to highlight or select images with novelty or 'interestingness'.

For the detection of novel scenes, a straightforward idea would be employing traditional statistical methods and machine learning methods [CBK09], and applying them to low-level visual features extracted from images. This is however rather questionable, since different objects may produce quite similar visual features and the ambiguity, long revealed by content-based image retrieval studies, cannot be resolved unless semantic analysis is conducted and objects within a scene are recognized. Recent psychophysical studies have revealed the importance of high-level semantics in object recognition and scene interpretation. It has been shown that top-down facilitation in recognition is triggered by early information about an object, and also by contextual

associations between the object and others with which it typically appears [FAGB06, SU07]. In light of these findings, it is reasonable to model the novelty of image scenes based on semantics within the scene.

In this paper, we consider the problem of detecting novelty in wildlife scenes. A computer system for this purpose would not only require the abilities of image analysis and object classification, but also the modelling of image semantics. For wildlife images, normal expectation is that wildlife animals reside in their habitats. So dolphins are usually in water, and zebras have land in the background. When a zebra appears with a dolphin in water, a novelty or anomaly can be detected. On the other hand, an image with similar objects may be classified into the same category; however, the different spatial context of the objects may create novelty. For example the two images in Fig. 1 have a similar occurrence of object classes ('elephant' and 'sky'), but the co-occurrence contexts are different. So Fig. 1(b) is treated as novel, because elephants are normally under the sky but not above it. We therefore propose a computational framework that models the semantic context, where novelty detection can be conducted based on comparing the co-occurrence of semantic labels.



(a) A 'normal' image

(b) A 'novel' image

Fig. 1. Images with similar objects but different context.

In the remainder of the paper, we first briefly review some relevant work in Section 2. Our computational framework and the novelty detection algorithm are introduced in Section 3, and the experiment results are presented in Section 4, followed by a conclusion and a discussion on future work.

2. RELATED WORK

Novelty detection has received most interest in textual information retrieval tasks [GDH04, LC08]. Novelty detection in video sequences has also been studied with the growing demand in surveillance applications [PLL07, Kha10]. In the area of wildlife surveillance [Con07], image analysis techniques have not found much utilization yet. Our approach differs from these in that we do not consider motion and trajectory information acquirable from videos, but rather concentrate on conducting semantic and statistical analysis of static scenes.

Closely related to novelty detection in images is that of interestingness discrimination and also image categorization. Even though human studies on interestingness discrimination has been carried out [KBCK08], relevant properties of images that are helpful for computational modelling in interestingness remains a challenge. Semantic representations of images have been explored for the tasks of content-based image retrieval and scene classification [WLW01, FFP05, JLZ+03]. The general approach is to identify semantics as the set of objects that appear in the image, and the scene is described by statistical modelling on the semantic objects [BNJ03, BZM06, FFFPZ05, LSFF09]. The relevant models are mainly applied in scene classification or understanding, but do not include novelty detection.

From a statistical point view, novelty detection is basically the same as anomaly detection. Many statistical and machine learning methods [MS03, CBK09] have been proposed and applied in various areas such as mechanics [MPW01] and network intrusion detection [PP07]. In this paper we follow the idea of distance

modelling and thresholding in deciding outliers [MPW01], but the statistical modelling is based on the co-occurrence of semantic labels.

3. COMPUTATIONAL FRAMEWORK

To achieve the goal of detecting novel or interesting images from a wildlife image collection, we propose a framework as shown in Fig. 2. The input images are first segmented into homogeneous segments, after which colour and texture features are extracted from image blocks within the image segments. These blocks can then be classified using classifiers trained on the visual features. To achieve the goal of novelty detection, we construct co-occurrence matrices of block labels and train a number of one-class classifiers, each modelled for a scene category. If a scene is rejected by all these one-class classifiers, then it is considered as novel or interesting. Hence, by going through the same process, both scene classification and novelty detection tasks can be accomplished.

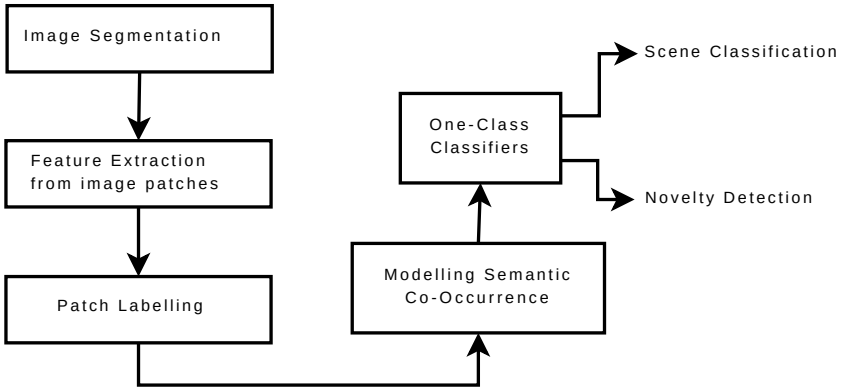


Fig. 2. The computational framework for image novelty detection and scene classification.

3.1. Image segmentation

A scene usually contains multiple objects of different visual characteristics. By segmenting an image into homogeneous regions, it facilitates detection or classification of these objects. We use the JSEG algorithm [DM01] for this purpose. In JSEG, colours in the image are first quantized to several representing classes that are used to separate the regions in the image. Image pixel colours are then replaced by their corresponding colour class labels to form a class-map of the image. In order to get good segmentation, the high and low values correspond to possible boundaries and interiors of colour-texture regions are adjusted until the object and background can be differentiated. Next, those segments that exceed a threshold will be taken out and stored as individual segment images respectively. Those segments will then be classified manually as the ground truth for training classifiers. Each segment image will be tiled into blocks of size 32×32. The image blocks that fall out of the segment edges will be ignored. For semantic analysis of new images, we have found it is more effective to train classifiers on segment blocks instead of on segment images directly.

3.2. Feature extraction

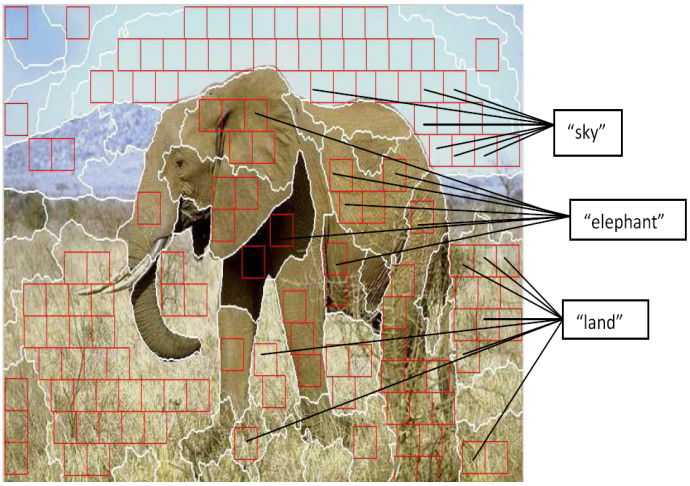
Visual features are then extracted from image segment blocks. First we employ the LUV colour histogram to encode the colour information of image blocks. Colour histograms are found to be robust to resolution and rotation changes. The LUV colour space is adopted because it models human's perception of colour similarity

very well, and it is also machine independent [MZ03]. Each channel is quantized with the same interval, thus L channel has 20-bins, U 70-bins and V 52-bins respectively. The standard deviation of the LUV histogram values is also calculated. The LUV histogram feature therefore has taken 143 dimensions.

Texture features extracted from the image blocks are also included as local features. We compared Edge Histogram Descriptor and Gabor filtering features [MOVY01] with the Haralick features [HSD73], and the latter performed the best in our test and are therefore adopted in further experiments. The Haralick texture features consist of a few statistical properties based on the gray scale co-occurrence matrix. The image block is first converted to gray scale. The co-occurrence matrix is a two-dimensional histogram composed from the pairwise statistics of gray scale co-occurrence among adjacent pixels. Four orientations are considered, each giving a co-occurrence matrix for an image block. A total of 13 statistical measures can be calculated for each co-occurrence matrix. The mean and deviation values of each of these 13 measures over the four orientations, form a feature vector of 26 dimensions.

Finally, the colour and texture features are concatenated together, giving a feature vector of 169 dimensions to represent an image block. Through manual labeling of image segments, semantic ground truth is assigned to the training images. The image blocks inherit semantic labels from their corresponding segments. Their feature codes along with the relevant class labels are used to train object classifiers.

3.3. Modelling the semantic context



class	coast	land	sky	water	dolphin	elephant	penguin	zebra
coast	0	0	0	0	0	0	0	0
land	0	1	0	0	0	1	0	0
sky	0	1	1	0	0	1	0	0
water	0	0	0	0	0	0	0	0
dolphin	0	0	0	0	0	0	0	0
elephant	0	1	1	0	0	1	0	0
penguin	0	0	0	0	0	0	0	0
zebra	0	0	0	0	0	0	0	0

Fig. 3. Image blocks in an 'elephant' image and its corresponding co-occurrence matrix.

To model the semantic context within a scene, we further generate a *block label co-occurrence matrix*

(BLCM). First the labels of all image blocks should be obtained - either by manual annotation for training images, or automatic classification when testing. Then the image blocks are scanned from left to right and top to bottom, and the co-occurrence of labels for blocks within a distance threshold R is collected. We set $R = 200$ in our experiment. The co-occurrence statistics is gathered across the entire image and normalized by the total number of image blocks. The variation on the object sizes does affect the matrix values of BLCM, so to reduce this effect the BLCM is binarized. Fig. 3 shows an 'elephant' image as an example, with blocks labels and the relevant binary BLCM shown below. There are 8 object labels: 'coast', 'land', 'sky', 'water', 'dolphin', 'elephant', 'penguin', and 'zebra', giving an 8×8 BLCM. Since the scanning of image blocks is directional, the BLCM is consequently asymmetric. This is however a desirable feature as we intend to keep the spatial location information in the semantic context representation. For example, a scene context of 'sky'→'land' is normal (scanning top-down), while a 'land'→'sky' BLCM entry may suggest some novelty. On the other hand, images with similar characteristics, i.e, with similar types of foreground and background settings will produce quite similar BLCM features, which is helpful for scene classification. In images with salient foreground objects, normally the BLCM is quite sparse. We concatenate the BLCM matrix rows into a 1-D vector, and principle component analysis (PCA) can then be used for dimension reduction. The dimension-reduced BLCM vectors are then ready to be further processed by classifiers.

3.4. Building classifiers

The classification task is two-fold. Given an image, if it is of an ordinary scene, the classifiers need to classify it into the right scene category; otherwise, novelty should be reported. This is however not a typical multi-class classification scenario, since there is no restriction of the appearance of novel images, and it is usually hard to find sufficient training data especially for the 'novelty' class. Hence we resort to one-class classification. To build a one-class classifier for each of the scene types, the classifiers need only normally labelled images for training. Testing images are assessed by calculating their feature vectors' distance to the trained one-class centers. If the distance is greater than a defined threshold for that class, it is rejected by that one-class classifier. A data instance rejected by all one-class classifiers is then reported as an outlier. Note that this approach carries out scene classification and novelty detection at the same time.

Our algorithm, called multiple one-class classification with distance thresholding (MOC-DT), is summarized as follows:

1. During training, for images of each image scene type, extract their BLCM code and calculate the mean, denoted as μ_i, of the image group. Here $i = 1, 2, ..., N$, N is the number of scene types.
2. Calculate the distance towards μ_i for all instances of the same scene and obtain the distance mean m_i, and the standard deviation, σ_i;
3. Set the outlier distance threshold as $T_i = m_i + 0.5\sigma_i$ for the scene type;
4. Given a query image, calculate its BLCM code and the distance towards each training model, d_i;
5. If $d_i < \min(T_i)$, assign the scene label k to the image, $k = \arg_i\min(T_i)$; otherwise (i.e., $d_i > T_i, \forall i$), label the image as 'novel'.

We use the Manhattan distance for the BLCM codes.

4. EXPERIMENTS AND RESULTS

4.1. Experiment settings

To validate the proposed approach, we conduct a few experiments using wildlife images and have the performance on scene classification and novelty detection measured.

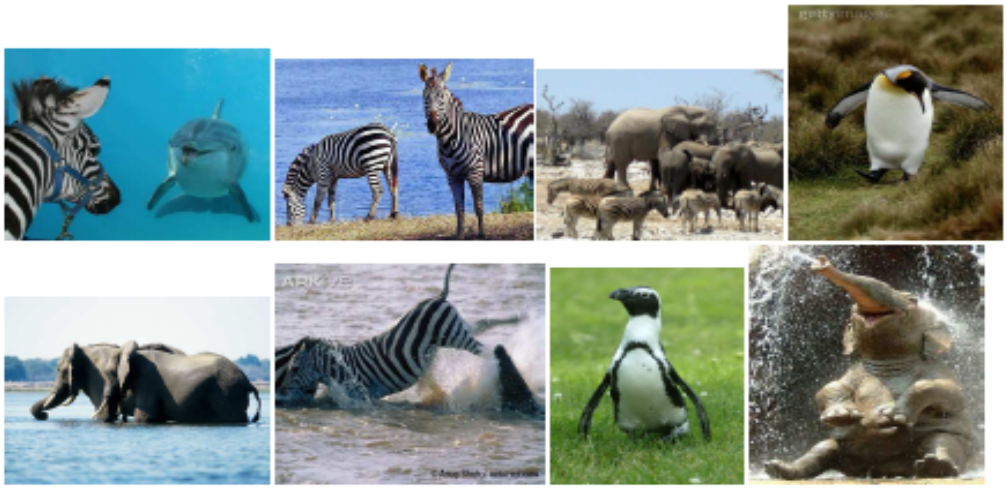
In training phase, 172 images are taken from ImageNet [DDS+09] and from Google Image Search. These wildlife images usually feature one animal in scene and contain relatively simple semantic context, each belonging to one of the following 4 scene types (number of instances for each type bracked as follows): 'dolphin' (39), 'elephant' (43), 'penguin' (45) and 'zebra' (45). Image sizes vary from 200×154 to 1024×768. Some sample images are shown in Fig. 4. Four background classes are included: 'coast', 'land, 'sky' and 'water'. For block labelling we have therefore 8 classes, including both the background and animals in foreground. After segmentation, there are over 13,000 image blocks extracted and used to train the classifier for the eight object classes, using the feature schemes presented in Section 3. For testing, we have 12 'normal' images plus 27 'novel' images. Novel images basically carry some different semantic context as compared with normal ones. For instance, scenes with zebra and dolphin together, penguin and land, elephant and coast etc., are considered 'novel'. Sample testing images are shown as in Fig. 5.

The classification of image blocks achieves an accuracy rate of 91% using the k-nearest neighbour classifier. This we believe gives a sound basis for constructing semantic context from labelled blocks.

Fig. 4. Samples of training images from four scene classes: 'dolphin', 'elephant', 'penguin' and 'zebra'.

4.2. Results

The semantic context of image scenes are calculated, undergoing binarization and PCA. We select the first 16 principal components to feed into classification. To visualize the generated BLCM codes for different scene types, a 2-D projection is produced using the first two principal components of the data, as shown in Fig. 6. Note that the projected data display a clustered structure with good separability. The 'elephant' and 'zebra'



(a) 'Novel' images

(b) 'Normal' images

Fig. 5. Sample images used for testing.

data points are quite close to each other, probably due to the fact that the relevant images have very similar background (i.e., mainly 'land').

First, we assess the performance of BLCM for scene classification using 10-fold cross validation on the training dataset. Two classifiers are employed: k-Nearest Neighbours (k-NN) and Random Forest (RF), both validated on two feature schemes: global scene descriptor (GSD) and BLCM. The GSD uses the same low-level features as used to construct BLCM: LUV colour histograms and Haralick texture features, but these are extracted globally (rather than locally in BLCM). The results, presented in Table 1, show that the binary BLCM with or without PCA achieves much higher accuracy in scene classification compared with using visual features directly.

For novelty detection, the testing images are processed by the proposed MOC-DT algorithm. The results of their respective distance towards each training model were shown in Fig. 7. It can be seen that the error instances as instance number 4, 5, 16 and 35. Some of these are shown in Fig. 8. The strong context contributed by abnormal background caused these errors.

The proposed MOC-DT algorithm is compared with the probability density based one-class classier (PDOC) [HFW08]. Each class model is treated as target class while the rest which are not classified into any particular class are considered as outliers. In other words, PDOC and MOC-DT are compared using the same treatment on the BLCM data, but their difference is only on the one-class classification algorithm. We also compare the algorithms' performance using BLCM and binarized BLCM with or without PCA. The results are shown in Table 2, reporting precision, recall and the F-score. We can see that BLCM works better

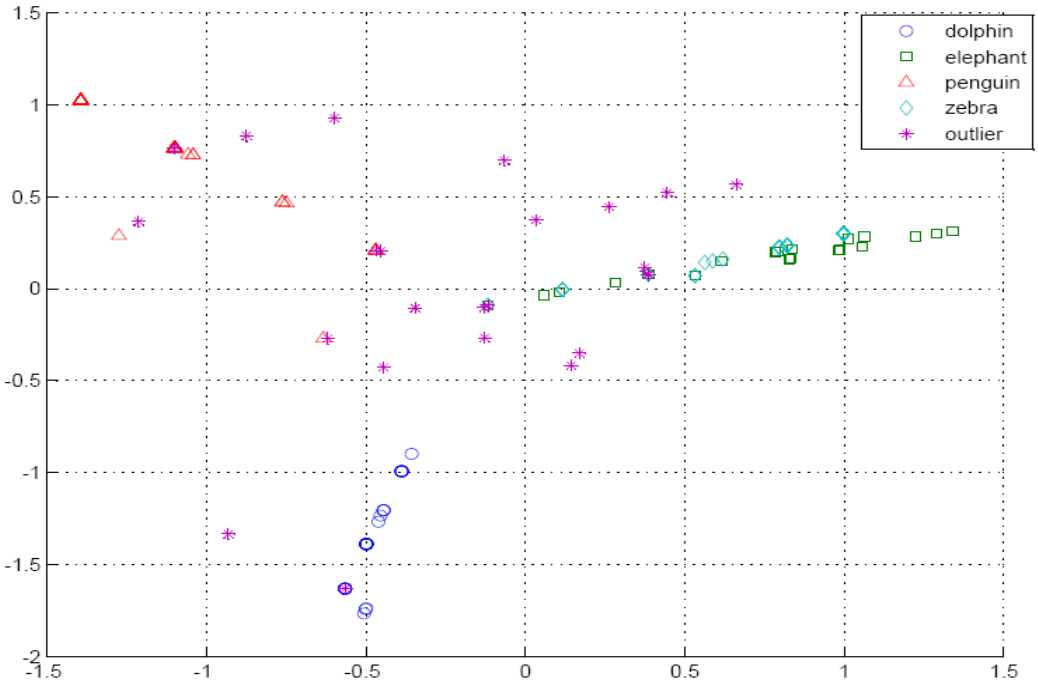


Fig. 6. The 2-D projection of BLCM data for normal and novel images. Normal images are marked by their classes and novel images as 'outlier'.

through binarization, and the best result is obtained using MOC-DT with binary BLCM plus PCA. PDOC performs worse than MOC-DT in all situations.

For the normal images (3 for each class) in the testing set, the MOC-DT algorithm manages classify all but one into the correct scene class, therefore giving an accuracy of 91.7%.

Table 1. Accuracy Results for Scene Classification

Feature schemes	Classifiers	
	RF	k-NN
GSD	66.86%	60.47%
Binary BLCM	94.77%	95.93%
Binary BLCM + PCA	**96.51%**	94.77%

5. CONCLUSION

Novelty detection of image scenes can be better achieved by working with semantic context modelling. In this paper we have proposed a simple but effective computational framework that conducts semantic context modelling and employs multiple one-class classifiers to carry out both scene classification and novelty detection. Our experiments on a set of images of four scene classes have given some promising results. Especially the proposed MOC-DT algorithm performs favourably compared with other algorithms. In the future we will take more object/scene categories into consideration and conduct experiments in larger scales.

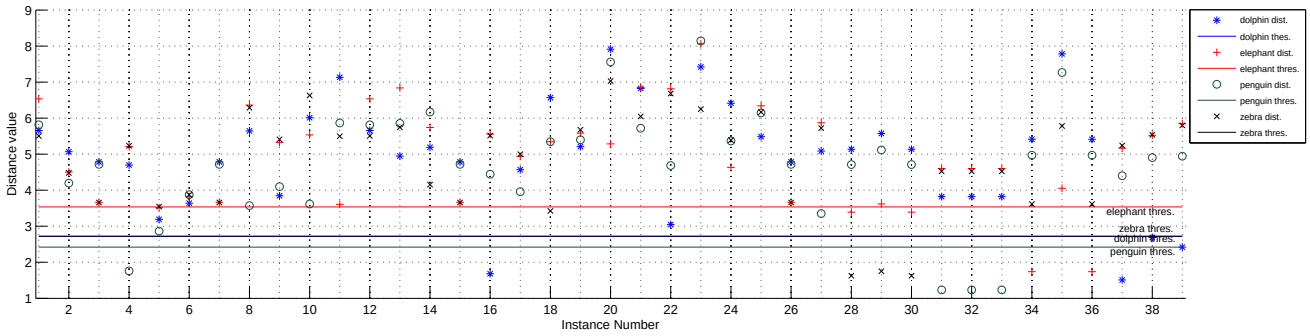


Fig. 7. Test instances with their respective distances towards each class model. Instances 1-22 belong to 'novel' classes, 23-39 are 'normal' images.

(a) Instance #4 (b) Instance #5 (c) Instance #16

Fig. 8. 'Novel' images misclassified as 'normal'.

On the other hand, the novelty or 'interestingness' of a scene is tightly coupled with visual experience and therefore can be dynamic and even subjective to the viewers. It is our intention to investigate using incremental models and algorithms to model the semantic context and its classification, and also incorporating user feedback to fine-tune these incremental models.

6. REFERENCES

[Ber60] D.E. Berlyne. *Stimulus Selection and Conflict*. McGraw-Hill Book Company, 1960.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Table 2. Novelty detection results obtained by MOC-DT and PDOC using BLCM or binarized BLCM (B-BLCM) without or without PCA.

Feature schemes	Classifiers					
	MOC-DT			PDOC		
	Pre.	Rec.	F-score	Pre	Rec	F-score
BLCM	0.95	0.74	0.83	1	0.30	0.46
BLCM+PCA	0.89	0.30	0.44	1	0.11	0.20
B-BLCM	0.96	0.82	0.88	1	0.33	0.50
B-BLCM+PCA	0.96	0.89	**0.92**	1	0.07	0.14

[BZM06] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In *Proceedings of the European Conference on Computer Vision*, pages 517–530, 2006.

[CBK09] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41:1–58, 2009.

[Con07] Christine Connolly. Wildlife-spotting robots. *Sensor Reviews*, 27:282–287, 2007.

[DDS+09] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Los Alamitos, CA, USA, 2009. IEEE Computer Society.

[DM01] Y. Deng, , and B.S. Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):800–810, Aug 2001.

[FAGB06] Mark J. Fenske, Elissa Aminoff, Nurit Gronau, and Moshe Bar. Top-down facilitation of visual object recognition: object-based and context-based contributions. *Progress in Brain Research*, 155:3–21, 2006.

[FFFPZ05] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *International Conference on Computer Vision*, volume 2, pages 1816–1823, 2005.

[FFP05] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scence categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 524–531, 2005.

[GDH04] Evgeniy Gabrilovich, Susan Dumais, and Eric Horvitz. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In *WWW'04: Proceedings of the 13th international conference on World Wide Web*, pages 482–490, New York, NY, USA, 2004. ACM.

[HFW08] K. Hempstalk, E. Frank, and I.H. Witten. One-class classification by combining density and class probability estimation. In *Proc. ECML/PKDD'08*, volume 5211 of *Lecture Notes in Computer Science*, pages 505–519, Berlin, September 2008. Springer.

[HSD73] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3:610–621, 1973.

[JLZ+03] F. Jing, M. Li, L. Zhang, H. Zhang, and B. Zhang. Learning in region-based image retrieval. In *International Conference on Image and Video Retrieval, Urbana-Champaign, Illinois*, 2003.

[KBCK08] H. Katti, K.Y. Bin, T.S. Chua, and M. Kankanhalli. Pre-attentive discrimination of interestingness in images. In *International Conference of Multimedia and Expo (ICME), Hannover, Germany*, June 23-26, 2008.

[Kha10] Shehzad Khalid. Motion-based behaviour learning, profiling and classification in the presence of anomalies. *Pattern Recognition*, 43(1):173 – 186, 2010.

[LC08] Xiaoyan Li and W. Bruce Croft. An information-pattern-based approach to novelty detection. *Information Processing and Management*, 44(3):1159 – 1188, 2008.

[LSFF09] L-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding:classification, annotation and segmentation in an automatic framework. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 2036–2043, 2009.

[MOVY01] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *Circuits and Systems for Video Technology, IEEE Transactions on*, 11(6):703–715, 2001.

[MPW01] G. Manson, G. Pierce, and K. Worden. On the long-term stability of normal condition for damage detection in a composite panel. *Key Engineering Materials*, 204-205:359–370, 2001.

[MS03] Markos Markou and Sameer Singh. Novelty detection: a review—part 1: statistical approaches. *Signal Process.*, 83(12):2481–2497, 2003.

[MZ03] Yu-Fei Ma and Hong-Jiang Zhang. Contrast-based image attention analysis by using fuzzy growing. In *Multimedia'03: Proceedings of the 11th ACM International Conference on Multimedia*, pages 374–381, New York, NY, USA, 2003. ACM.

[PLL07] Dragoljub Pokrajac, Aleksandar Lazarevic, and Longin Jan Latecki. Incremental local outlier detection for data streams. In *Proceedings of IEEE Symposium on Computational Intelligence and Data Mining*, pages 504–515, 2007.

[PP07] Animesh Patcha and Jung-Min Park. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, 51(12):3448 – 3470, 2007.

[Sil06] Paul J. Silvia. *Exploring the Psychology of Interest*, volume 56. Oxford University Press, 2006.

[SU07] J. A. Stirk and G. Underwood. Low-level visual saliency does not predict change detection in natural scences. *Journal of Vision*, 7(10):3:1–10, 2007.

[WLW01] J. Wang, J. Li, and G. Wiederhold. Simplicity: semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, 2001.